

Claims

We claim:

1. For use with a search engine that processes user queries, a system that locates documents containing words corresponding to a user query comprising;
an infrequent word identifier that identifies infrequent words that occur in less than a threshold number of documents;
a frequent word index that maps the location of documents that contain words that occur in more than the threshold number of documents;
an infrequent word index, maintained separately from the frequent word index, that maps the location of documents that contain infrequent words;
an index scanning component that, in response to a query containing an infrequent word, scans the infrequent word index to find the location of documents containing the infrequent word.
2. The system of claim 1 wherein the frequent word index is stored by document.
3. The system of claim 1 wherein the frequent word index is partitioned by document.
4. The system of claim 3 wherein the frequent word index is distributed across multiple computing systems.
5. The system of claim 1 wherein the infrequent word index is stored by document.
6. The system of claim 1 wherein the infrequent word index is partitioned by document.

7. The system of claim 6 wherein the infrequent word index is distributed across multiple computing computer systems.
8. The system of claim 1 wherein the infrequent word index is stored by word.
9. The system of claim 1 wherein the infrequent word index is partitioned by word.
10. The system of claim 9 wherein the infrequent word index is stored on a single computing computer system.
11. The system of claim 10 wherein the index scanning component, in response to a user query containing an infrequent word, retrieves document locations for documents having the infrequent word from the infrequent word index and transmits the retrieved document locations to computer systems containing frequent word indexes for the retrieved documents.
12. The system of claim 1 including an index cache associated with the infrequent word index that stores document locations for recently queried infrequent words.

13. For use with a search engine that processes user queries, a method that searches a set of documents for documents containing terms found in a user query comprising:

scanning the set of documents and gathering infrequent words that occur a number of times that is less than a threshold amount;

constructing an infrequent word index that maps infrequent words to locations of documents that contain the words;

constructing a frequent word index, separately maintained from the infrequent word index, that maps frequent words that occur a number of times that is greater than the threshold amount to locations of documents that contain the words; and

examining the terms in the user query to identify any terms are infrequent words; and

searching the infrequent word index for the terms that are identified as infrequent words.

14. The method of claim 13 comprising storing the infrequent word index in a dedicated computer system.

15. The method of claim 13 comprising storing the infrequent word index in dedicated partitions on computer systems that also store the frequent word index.

16. The method of claim 13 comprising storing the infrequent index by word.

17. The method of claim 13 comprising storing the infrequent index by document.

18. A computer readable medium comprising computer-executable instructions for performing the method of claim 13.

19. For use with a search engine that processes user queries, a computer readable medium comprising computer-executable instructions for locating documents containing words corresponding to a user query by:

identifying infrequent words that occur in less than a threshold number of documents;

mapping the location of documents that contain words that occur in more than the threshold number of documents in a frequent word index;

maintaining, separately from the frequent word index, an infrequent word index that maps the location of documents that contain infrequent words;

in response to a query containing an infrequent word, scanning the infrequent word index to find the location of documents containing the infrequent word.

20. The computer readable medium of claim 19 wherein the infrequent word index is stored by document.

21. The computer readable medium of claim 19 wherein the infrequent word index is partitioned by document.

22. The computer readable medium of claim 19 wherein the infrequent word index is distributed across multiple computing computer systems.

23. The system of claim 1 wherein the infrequent word index is stored by word.

24. The computer readable medium of claim 19 wherein the infrequent word index is partitioned by word.

25. The computer readable medium of claim 19 wherein the infrequent word index is stored on a single computing computer system.

26. The computer readable medium of claim 19 including creating an index cache associated with the infrequent word index that stores document locations for recently queried infrequent words.

27. For use with a search engine that processes user queries, an apparatus for searching set of documents for documents containing terms found in a user query comprising:

means for scanning the set of documents and gathering infrequent words that occur a number of times that is less than a threshold amount;

means for constructing an infrequent word index that maps infrequent words to locations of documents that contain the words;

means for constructing a frequent word index, separately maintained from the infrequent word index, that maps frequent words that occur a number of times that is greater than the threshold amount to locations of documents that contain the words; and

means for examining the terms in the user query to identify any terms are infrequent words; and

means for searching the infrequent word index for the terms that are identified as infrequent words.